THE CONDITION NUMBER OF A CLASS OF RAYLEIGH-RITZ-GALERKIN MATRICES[1]

BY MARTIN H. SCHULTZ

Communicated by Eugene Isaacson, February 5, 1970

The purpose of this note is to study the Euclidean condition number of the matrix resulting from using the well-known Rayleigh-Ritz-Galerkin method with finite dimensional subspaces of polynomial spline functions to approximate the solution of a linear, self-adjoint, two-point boundary value problem. Roughly speaking, we consider a model class of such problems of order $2n$ and determine *upper bounds*, of the form of a constant times the norm of the partition associated with the spline subspace to the $-2n$th power, for the Euclidean condition number of the associated matrix.[2]

The class of problems we are considering is defined by

$$(1) \qquad L[u] \equiv \sum_{j=0}^{n} (-1)^j D^j[p_j(x) D^j u(x)] = f(x), \qquad -\infty < x < \infty,$$

subject to the boundary conditions

$$(2) \qquad \lim_{x \to \infty} D^k u(x) = \lim_{x \to -\infty} D^k u(x) = 0, \qquad 0 \leqq k \leqq n-1,$$

where $D \equiv d/dx$.

Let $H_0^{n,2}$ be the completion of the real-valued functions, $h(x)$, in $C_0^\infty(-\infty, \infty)$, i.e., the completion of the infinitely differentiable, real-valued functions with compact support, with respect to the Sobolev norm

$$\|h\|_n \equiv \left(\int_{-\infty}^{\infty} [D^n h(x)]^2 dx \right)^{1/2}.$$

Moreover, we assume that $p_j(x) \in L^\infty(-\infty, \infty)$ and are real-valued for $0 \leqq j \leqq n$, $f(x) \in L^2(-\infty, \infty)$ and is real-valued, and that

AMS Subject Classifications. Primary 6566, 6562, 6550.

Key Words and Phrases. Rayleigh, Ritz, Galerkin, matrices, splines, condition number.

[1] This research was supported in part by the National Science Foundation, GP 11326.

[2] After sending this manuscript to the editor we learned that G. Fix and G. Strang have obtained analogous results for the case of uniform partitions by means of Fourier transform techniques.

$$K_1\|h\|_n^2 \leqq \int_{-\infty}^{\infty} \sum_{j=0}^{n} p_j(x)[D^j h(x)]^2 dx \leqq K_2\|h\|_n^2, \tag{3}$$

for all $h \in H_0^{n,2}$. We remark that with these hypotheses, the problem (1)–(2) has a unique generalized solution, cf. [2].

To approximate the solution $u(x)$, we define a and b large enough in absolute value so that the solution, $u(x)$, and its derivatives are close to zero outside of $[a, b]$ and let $\{B_i(x)\}_{i=1}^N \subset H_0^{n,2}$ be N given linearly independent functions such that supp $B_i \subset (a, b)$, $1 \leqq i \leqq N$. We seek an approximation to $u(x)$ of the form $v(x) = \sum_{i=1}^N \beta_i B_i(x)$, where we determine the coefficients $\{\beta_i\}_{i=1}^N$ by the Rayleigh-Ritz-Galerkin method. For example, in the Galerkin method we demand that the residual $L[\sum_{i=1}^N \beta_i B_i(x)] - f(x)$ be orthogonal in $L^2(a, b)$ to the basis functions $\{B_i(x)\}_{i=1}^N$, i.e.,

$$\int_a^b \sum_{j=0}^{n} p_j(x)\left[\sum_{i=1}^{N} \beta_i D^j B_i(x)\right] D^j B_k(x)dx = \int_a^b f(x)B_k(x)dx, \tag{4}$$

for $1 \leqq k \leqq N$, where we have obtained the left-hand side by integrating by parts.

The system (4) may be rewritten in matrix form

$$A\beta = k, \tag{5}$$

where

$$A \equiv \left[\int_a^b \sum_{j=0}^{n} p_j(x) D^j B_i(x) D^j B_k(x)dx\right]$$

and $k \equiv [\int_a^b f(x)B_k(x)dx]$. Clearly, the matrix A is symmetric and positive definite. In fact, using (3) and the Rayleigh-Ritz inequality, cf. [3], we have that

$$\begin{aligned}
\beta^t A\beta &= \sum_{j=0}^{n} \int_a^b p_j(x)\left[\sum_{i=1}^{N} \beta_i D^j B_i(x)\right]^2 dx \\
&= \sum_{j=0}^{n} \int_{-\infty}^{\infty} p_j(x)\left[\sum_{i=1}^{N} \beta_i D^j B_i(x)\right]^2 dx \\
&\geqq K_1 \int_{-\infty}^{\infty}\left[\sum_{i=1}^{N} \beta_i D^n B_i(x)\right]^2 dx \\
&= K_1 \int_a^b \left[\sum_{i=1}^{N} \beta_i D^n B_i(x)\right]^2 dx \\
&\geqq K_1\left(\frac{\pi}{b-a}\right)^{2n} \int_a^b \left[\sum_{i=1}^{N} \beta_i B_i(x)\right]^2 dx > 0
\end{aligned}$$

unless $\beta = 0$, since the B_i's are linearly independent.

We now consider the special case of spline basis functions. Following a construction due to de Boor, cf. the fundamental [1], if d is a positive integer, a finite set of real numbers $\Delta: a = x_0 \leqq x_1 \leqq \cdots \leqq x_N \leqq x_{N+1} = b$ is said to be a $d+1$-*extended partition* of the interval $[a, b]$ if and only if $x_i < x_{i+d}$ for all $0 \leqq i \leqq N-d+1$, i.e. if f_i denotes the frequency with which x_i occurs in Δ, then $f_i \leqq d$ for all $0 \leqq i < N-d+1$. Let $I \equiv \{0 \leqq i \leqq N \mid x_i < x_{i+1}\}$, $M_i(x) \equiv (d+1)t(x_i, \cdots, x_{i+d+1}; x)$ be the $(d+1)$-times divided difference in y of the function $t(x, y) \equiv (y-x)_+^d$ based on the points $x_i, \cdots, x_{i+d+1}$, and

$$B_i(x) \equiv \frac{x_{i+d+1} - x_i}{d+1} M_i(x) \quad \text{for all } 0 \leqq i \leqq N-d.$$

It follows that $B_i(x) \geqq 0$ for all $-\infty < x < \infty$ with equality if and only if $x \notin (x_i, x_{i+d+1})$ for all $0 \leqq i \leqq N-d$, and $\sum_{i=0}^{N-d} B_i(x) \leqq 1$ for all $-\infty < x < \infty$. Let $S_0(d, \Delta)$ be the linear span of $\{B_i\}_{i=f_0}^{N-d-f_{N-d}}$. It is easy to see that if $s(x) \in S_0(d, \Delta)$, then $s(x)$ reduces to a polynomial of degree d on each of the subintervals $[x_i, x_{i+1}]$ for all $i \in I$ and has $d-f_i$ continuous derivatives in a neighborhood of x_i for all $1 \leqq i \leqq N$. Moreover, supp $s(x) \subset (a, b)$ and if $d-f_i \geqq n-1$ for all $f_0 \leqq i \leqq N-d-f_{N-d}$ then $S_0(d, \Delta) \subset H_0^{n,2}$. We consider a reordering of the basis functions such that $f_0 \equiv 1$ and $N-d-f_{N-d} \equiv M$ and hence $S_0(d, \Delta)$ is the span of $\{B_i\}_{i=1}^{M}$. Moreover, it follows from Corollary 1 of Theorem 3.1 of [1] that there exist positive constants, Q_d, depending only on d, such that

$$Q_d \|\beta\|_\infty \leqq \left\| \sum_{i=1}^{M} \beta_i B_i(x) \right\|_{L^\infty(a,b)} \leqq \|\beta\|_\infty, \tag{6}$$

where

$$\|\beta\|_\infty \equiv \max_{1 \leqq i \leqq M} |\beta_i|, \quad \text{for all } \beta \in R^M.$$

Finally, we define $\bar{\Delta} \equiv \max_{i \in I}(x_{i+1} - x_i)$ and $\Delta \equiv \min_{i \in I}(x_{i+1} - x_i)$. We now prove the main result of this paper.

THEOREM. *If* (3) *holds,* Δ *is a* d*-extended partition of* $[a, b]$, $d-f_i \geqq n-1$ *for all* $f_0 \leqq i \leqq N-d-f_{N-d}$, *and* A *is the matrix of the linear system given in* (4), *then*

$$\operatorname{cond}(A) \equiv \|A\|_2 \|A^{-1}\|_2 \leqq \frac{4^n(d+1)^2 d^{4n+1}(b-a)^{2n} K_2}{2K_1 Q_d^2 \pi^{2n-2}} \cdot \bar{\Delta}(\Delta)^{-1}(\Delta)^{-2n}, \tag{7}$$

where

$$\|A\|_2 \equiv \sup_{\beta \neq 0} \|A\beta\|_2/\|\beta\|_2 \quad \text{and} \quad \|\beta\|_2 \equiv \left(\sum_i |\beta_i|^2 \right)^{1/2}.$$

PROOF. Since A is a symmetric matrix, $\text{cond}(A) = \lambda^{-1}\Lambda$, where λ is the minimum eigenvalue of A and Λ is the maximum eigenvalue of A. Hence, it suffices to obtain upper bounds for λ^{-1} and Λ. From (3) and the Rayleigh-Ritz inequality and the boundary conditions we have

$$\begin{aligned}
\beta^t A \beta &= \sum_{j=0}^{n} \int_a^b \left[\sum_{i=1}^{M} \beta_i D^j B_i(x) \right]^2 dx \geqq K_1 \int_a^b \left[\sum_{i=1}^{M} \beta_i D^n B_i(x) \right]^2 dx \\
&\geqq K_1 \left(\frac{\pi}{b-a} \right)^{2n-2} \int_a^b \left[\sum_{i=1}^{M} \beta_i DB_i(x) \right]^2 dx \\
&\geqq K_1 \left(\frac{\pi}{b-a} \right)^{2n-2} \frac{2}{b-a} \left\| \sum_{i=1}^{M} \beta_i B_i(x) \right\|^2_{L^\infty(a,b)} \\
&\geqq \frac{2K_1 Q_d^2 \pi^{2n-2}}{(b-a)^{2n-1}} \|\beta\|_\infty^2 \geqq \frac{2K_1 Q_d^2 \pi^{2n-2}}{(d)(b-a)^{2n}} \Delta \|\beta\|_2^2,
\end{aligned}$$

since β has at most $(d)(b-a)/\Delta$ components. Thus,

$$\lambda \geqq \frac{2K_1 Q_d^2 \pi^{2n-1}}{(d)(b-a)^{2n}} \Delta. \tag{8}$$

Conversely, using the Markov inequality cf. [4], we have

$$\begin{aligned}
\beta^t A \beta &\leqq K_2 \int_a^b \left[\sum_{i=1}^{M} \beta_i D^n B_i(x) \right]^2 dx \\
&\leqq (d+1) K_2 \sum_{i=1}^{M} \beta_i^2 \int_a^b [D^n B_i(x)]^2 dx \\
&= (d+1) K_2 \sum_{i=1}^{M} \beta_i^2 \int_{x_i}^{x_{i+d+1}} [D^n B_i(x)]^2 dx \\
&\leqq (d+1) K_2 \sum_{i=1}^{M} \beta_i^2 \|D^n B_i(x)\|^2_{L^\infty(-\infty,\infty)} \int_{x_i}^{x_{i+d+1}} 1 dx \\
&\leqq \frac{(d+1)^2 K_2 (2d^2)^{2n} \bar{\Delta} \|\beta\|_{l^2}^2}{(\Delta)^{2n}}.
\end{aligned}$$

Thus, we have

$$\Lambda \leqq 4^n (d+1)^2 d^{4n} K_2 \bar{\Delta} (\Delta)^{-2n}. \tag{9}$$

Combining (8) and (9), we obtain the required result. Q.E.D.

COROLLARY. *If* (3) *holds, C is a collection of d-extended partitions,* Δ, *of* $[a, b]$, *such that* $d-f_i \geqq n-1$ *for all* $f_0 \leqq i \leqq N-d-f_{N-d}$ *for all* $\Delta \in C$, $\bar{\Delta}(\Delta)^{-1} \leqq \eta < \infty$ *for all* $\Delta \in C$, *and* $A(\Delta)$ *is the matrix of the linear system given in* (4) *for* $S_0(d, \Delta)$, *then there exists a positive constant, K, independent of* $\bar{\Delta}$, *such that*

$$\text{cond}(A(\Delta)) \leqq K(\bar{\Delta})^{-2n} \tag{10}$$

for all $\Delta \in C$.

We remark that the exponent in (10) is *independent* of d and that in the special case of $n=1$, (10) shows that the matrices A for spline subspaces are conditioned no worse than the analogous matrices obtained from the standard three point central difference approximation to (1).

REFERENCES

1. C. de Boor, *On uniform approximation by splines*, J. Approximation Theory **1** (1968), 219–235. MR **39** #1866.

2. J. Céa, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier (Grenoble) **14** (1964), fasc. 2, 345–444. MR **30** #5037.

3. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd. ed., Cambridge Univ. Press, Cambridge, 1952. MR **13**, 727.

4. J. Todd, (Editor) *Survey of numerical analysis*, McGraw-Hill, New York, 1962. MR **24** #B1271.

CALIFORNIA INSTITUTE OF TECHNOLOGY, PASADENA, CALIFORNIA 91109[3]

[3] The author's present address is Computer Science Department, Yale University.